

Econ 673: Microeconometrics

Chapter 3: Numerical Maximization

Fall 2008

Outline

- 1 Algorithms
 - Grid Searches
 - Iterative Algorithms
- 2 Convergence Criteria
- 3 Parameter Transformations

Readings

- Required:
 - Train, K., (2003), *Discrete Choice Methods with Simulation*, Cambridge, MA: Cambridge University Press, Ch. 8.
- Recommended:
 - Greene, W. H., *Econometric Analysis*, 6th edition, Upper Saddle River, New Jersey: Prentice-Hall, Inc., Appendix E.3.
 - Ruud, P., *An Introduction to Classical Econometric Theory*, New York: Oxford University Press, 2000, Ch. 16.

The Problem

Most of the estimation problems we consider take the form:

$$\text{Max}_{\theta} F(\theta) \quad (1)$$

Maximum Likelihood:

$$\begin{aligned} F(\theta) &= \frac{1}{N} LL(\mathbf{y}|\theta, \mathbf{X}) \\ &= \frac{1}{N} \sum_{i=1}^N \ln [P(y_i|\theta, \mathbf{x}_i)] \end{aligned}$$

Nonlinear Least Squares:

$$F(\theta) = -\frac{1}{N} \sum_{i=1}^N [y_i - f(\theta, \mathbf{x}_i)]^2 \quad (2)$$

Grid Searches

A simple grid search can be a reliable approach to finding maxima over a closed interval; i.e.,

$$\underset{\theta \in [a,b]}{\text{Max}} F(\theta) \quad (3)$$

- 1 Divide $[a, b]$ into finite subintervals

$$\{[\theta_0 = a, \theta_1], [\theta_1, \theta_2], \dots, [\theta_{K-1}, \theta_K = b]\} \quad (4)$$

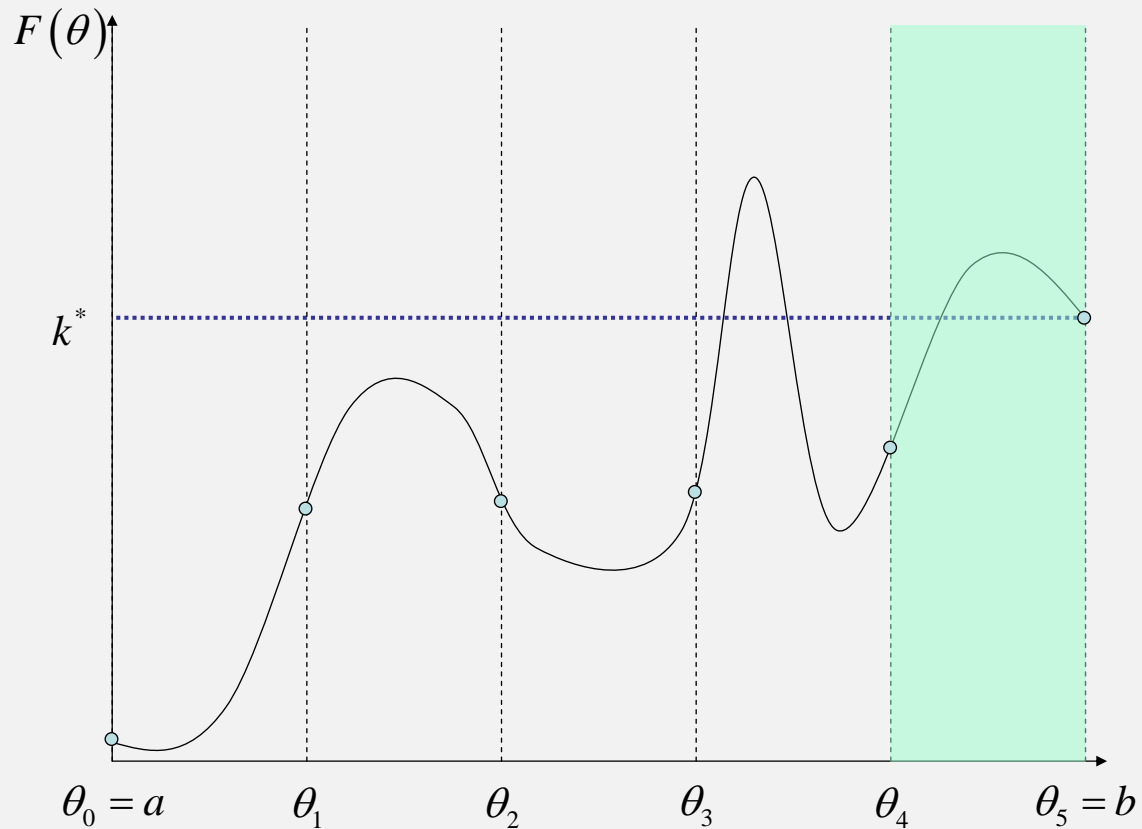
- 2 Compute $k^* = \underset{k}{\text{argMax}} F(\theta_k)$
- 3 Check for convergence. If not, set

$$a = \theta_{k^*-1}; \quad b = \theta_{k^*+1} \quad (5)$$

and return to step 1.

Grid Search (cont'd)

- Advantages:
 - Simple - works well for one-dimensional problems with a concave objective function and well defined boundaries
- Disadvantages:
 - For higher dimensional problems, can be very slow requiring a large number of grid points per iteration, though still useful for finding starting values
 - When the objective function is not concave, a coarse grid may miss global maximum



Iterative Algorithms

Most algorithms used for nonlinear optimizations are iterative

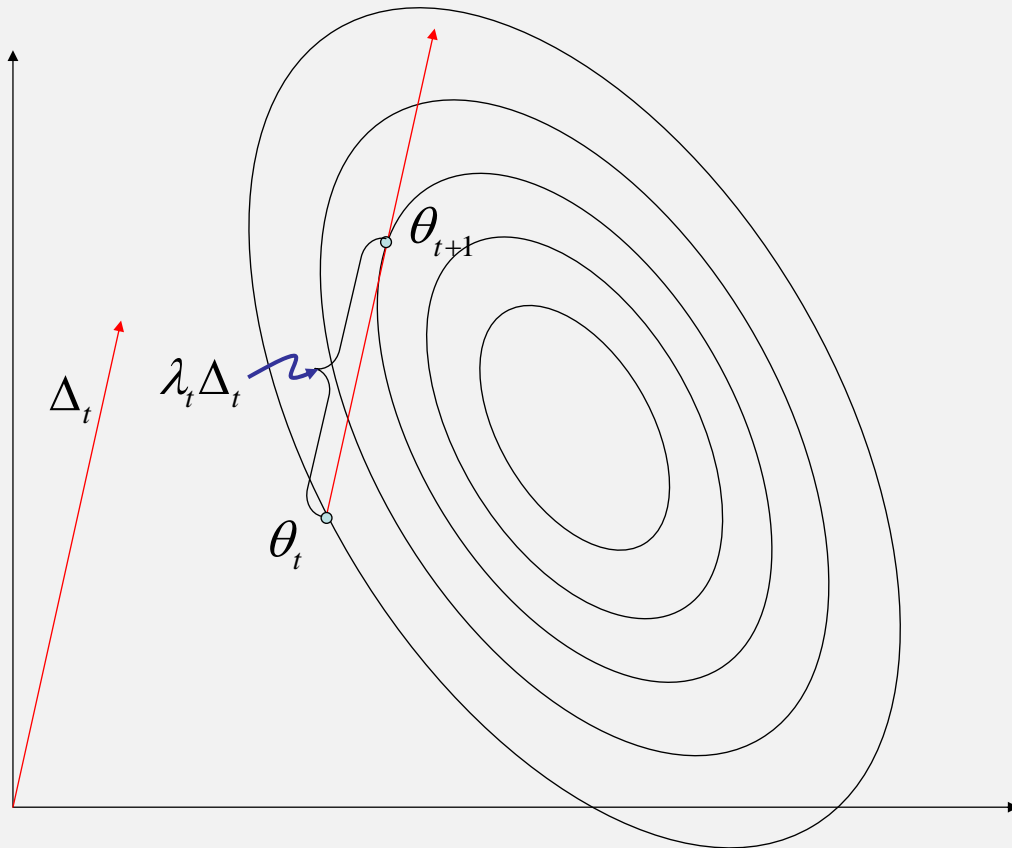
$$\theta_{t+1} = \theta_t + \lambda_t \Delta_t \quad (6)$$

where

Δ_t denotes the direction of change

$\lambda_t > 0$ denotes the step length

The differences in algorithms center around choices of Δ_t and λ_t .



Optimal Step Length

For a given Δ_t , one can optimize the step length λ_t

$$0 = \frac{\partial F(\theta_t + \lambda_t^* \Delta_t)}{\partial \lambda_t} = g(\theta_t + \lambda_t^* \Delta_t)' \Delta_t \quad (7)$$

where

$$g(\theta) \equiv \frac{\partial F(\theta)}{\partial \theta}. \quad (8)$$

However, usually informal searches are used.

Gradient Methods

Most common, with

$$\Delta_t = W_t g(\theta_t) \quad (9)$$

where W_t is positive definite. Rationale is based on first-order Taylor series approximation:

$$\begin{aligned} F(\theta_{t+1}) &= F(\theta_t + \lambda_t W_t g_t) \\ &\approx F(\theta_t) + \lambda_t g_t' W_t g_t \end{aligned}$$

So that

$$F(\theta_{t+1}) - F(\theta_t) \approx \lambda_t g_t' W_t g_t > 0$$

Steepest Ascent

Yields greatest increase in objective function *among all changes of the same length*.

$$W_t = I \Rightarrow \Delta_t = g_t \quad (10)$$

- Advantage: Optimal step length has closed form, with

$$\lambda_t^* = \frac{-g_t' g_t}{g_t' H_t g_t} \quad (11)$$

where

$$H_t \equiv \frac{\partial^2 F(\theta)}{\partial \theta \partial \theta'}. \quad (12)$$

- Disadvantage: Slow to converge
Does not exploit curvature of objective function

Newton-Raphson

Based on second-order Taylor series approximation

$$F(\theta_{t+1}) = F(\theta_t) + (\theta_{t+1} - \theta_t)' g_t + \frac{1}{2} (\theta_{t+1} - \theta_t)' H_t (\theta_{t+1} - \theta_t)$$

Maximizing with respect to θ_{t+1}

$$\begin{aligned} 0 &= \frac{\partial F(\theta_{t+1})}{\partial \theta_{t+1}} \\ &= g_t + H_t (\theta_{t+1} - \theta_t) \\ &\Rightarrow \\ \theta_{t+1}^* &= \theta_t - H_t^{-1} g_t \end{aligned}$$

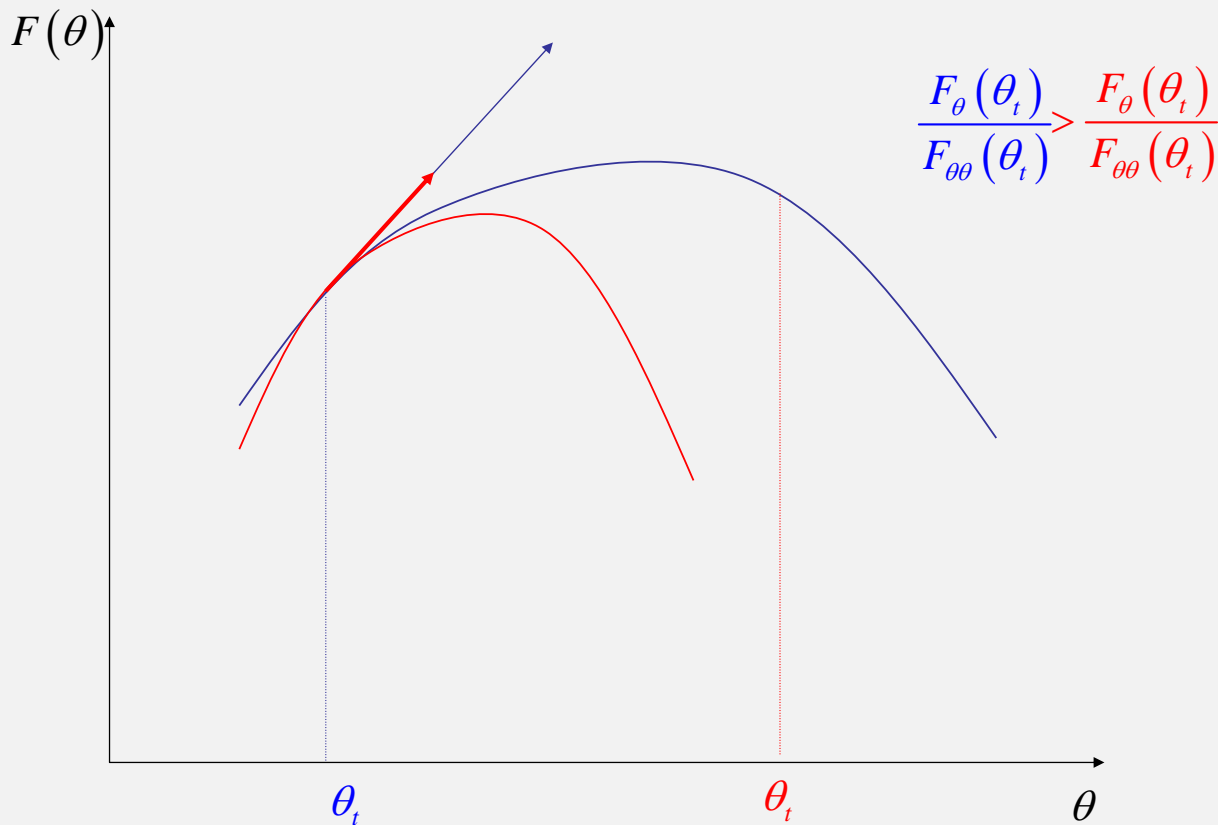
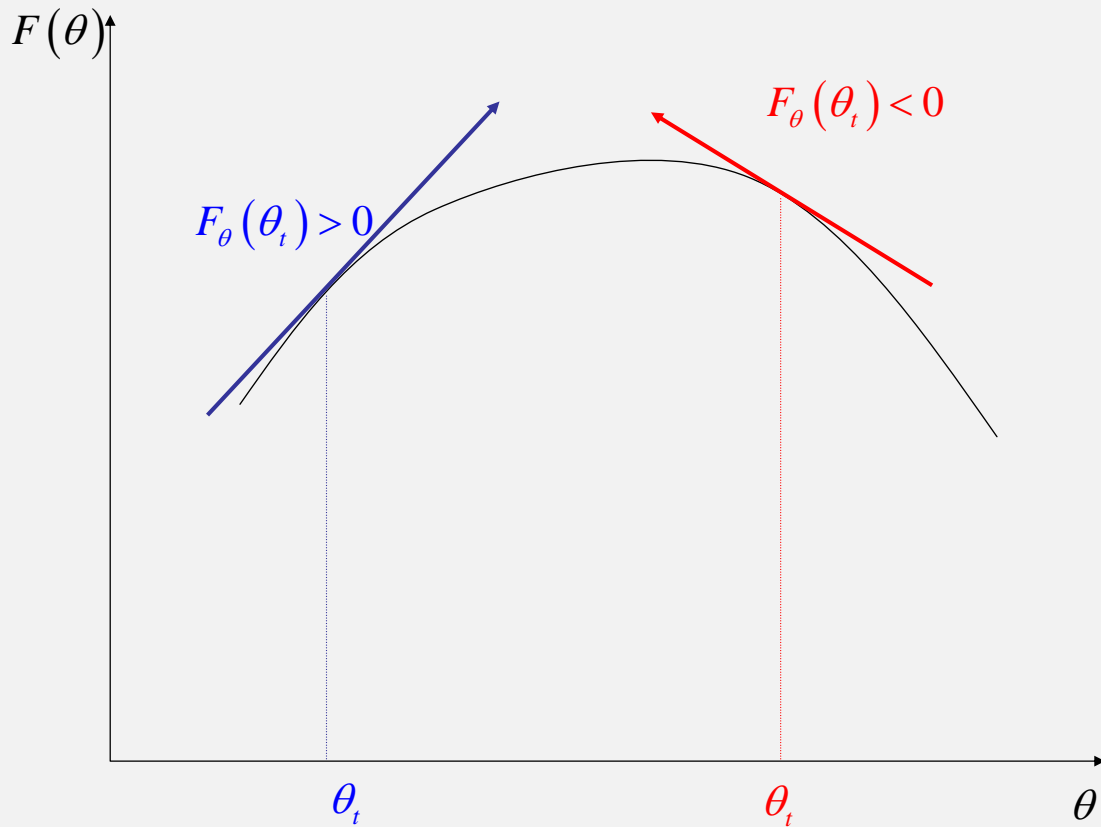
Basic NR uses $W_t = -H_t^{-1} \Rightarrow \Delta_t = -H_t^{-1} g_t$ and $\lambda_t = 1$

Intuition

Consider single parameter case, with

$$\theta_{t+1} - \theta_t = -H_t^{-1} g_t = \frac{F_\theta}{-F_{\theta\theta}} \begin{cases} > 0 & F_\theta > 0 \\ < 0 & F_\theta < 0 \end{cases}$$

If function, evaluated at last iteration is increasing, then the next iteration increases the parameter.



Step Length

- Most packages start with $\lambda_t = 1$
- If improvement in F is not found, then

$$\lambda_t = \left(\frac{1}{2}\right)^k ; k = 2, \dots, K \quad (13)$$

are tried sequentially. This is a low cost alternative to optimizing λ_t

- Some packages will also try increasing λ_t .

Relative Merits of Newton-Raphson

- Unlike Steepest Ascent, relies on *both* slope and curvature of objective function
- If objective function is quadratic, will converge in one iteration

Limitations:

- Computationally intensive, requiring Hessian
- When Hessian is not negative definite, increase is not guaranteed.

Quasi-Newton Modifications

- Use

$$W_{t+1} = W_t + E_t \quad (14)$$

where E_t and W_0 are both positive definite, insuring the W_t is positive definite for all t .

- Two prominent examples are
 - Davidson Fletcher-Powell (DFP) and
 - Broyden-Fletcher-Goldfarb-Shanno (BFGS)
- Many packages use DFP or BFGS as default algorithms
- Train (2003) notes that both are based on providing arc Hessians, thus using multiple points to determine curvature.

Maximum Likelihood

- In the case of MLE, the gradient of the objective function becomes the sum of scores

$$\begin{aligned} g(\theta) &= \frac{\partial LL(\mathbf{y}|\theta, \mathbf{X})}{\partial \theta} \\ &= \frac{1}{N} \sum_{i=1}^N \frac{\partial \ln [P_i(\mathbf{y}_i|\theta, \mathbf{X}_i)]}{\partial \theta} \\ &= \frac{1}{N} \sum_{i=1}^N s_i(\mathbf{y}_i|\theta, \mathbf{X}_i) \end{aligned}$$

- The Information Identity states that, for a correctly specified model and at the true parameters, the covariance of first derivatives equals the negative of the average of second derivatives.

Berndt-Hall-Hall-Hausman (BHHH)

- Uses:

$$W_t = g_t g_t' = \frac{1}{N} \sum_{i=1}^N s_i(\mathbf{y}_i | \theta, \mathbf{X}_i) s_i(\mathbf{y}_i | \theta, \mathbf{X}_i)' \quad (15)$$

- If the average score is zero, then W_t is the sample variance of the scores.
- If W_t is large, this suggests lots of variability among observations (i.e., lots of information) and peaked likelihood function and one should take small steps.

Relatives Merits of BHHH

Advantages:

- W_t is guaranteed to be positive definite.
- The computation of W_t requires only gradients.

Disadvantages:

- Information identity only holds at true parameter values
- Train (2003) suggests using mean deleted score in computing W_t .

Simple Convergence Criteria

A tempting convergence criteria is

$$|\theta_{t+1} - \theta_t| \leq \kappa \quad (16)$$

where κ is the established *tolerance* level. This is *not* a good choice, as it is sensitive to

- Small step lengths
- Scaling

Alternative Criteria

- An improved criteria, avoiding the scaling issue would be

$$\left| \frac{\theta_{t+1} - \theta_t}{\theta_t} \right| \leq \kappa \quad (17)$$

- The most commonly used convergence statistic is:

$$-\mathbf{g}_t' \mathbf{H}_t^{-1} \mathbf{g}_t \leq \kappa \quad (18)$$

Local versus Global Maxima

- While a number of the models we will consider have globally concave likelihood functions, other will not.
- Always try a variety of starting values to determine if the results are sensitive to starting values and, hence, the presence of local maxima.

Parameter Transformations

Parameter Transformations

- Rescaling of variables should be used to yield parameters of the same order of magnitude.
- Parameter Transformations can be used to impose prior restrictions:
 - $\theta > 0$ Substitute $\theta = \exp(\alpha)$, estimating α
 - $\theta \geq 0$ Substitute $\theta = \alpha^2$, estimating α
 - $\theta \in (0, 1)$ Substitute $\theta = \frac{1}{1 + \exp(\alpha)}$, estimating α
 - $\theta \in (-1, 1)$ Substitute $\theta = \frac{1 - \exp(\alpha)}{1 + \exp(\alpha)}$, estimating α

Concentrating the Likelihood Function

- There are many problems in which sequential solution of the ML problem is convenient, with

$$\theta_2^* = t(\theta_1) = \underset{\theta_2}{\text{Max}} F(\theta_1, \theta_2) \quad (19)$$

- The objective function is then *concentrated* as

$$F^*(\theta_1, \theta_2) = F(\theta_1, t(\theta_1)) = F_c(\theta_1) \quad (20)$$

- The unrestricted maximization is then solved by

$$\begin{aligned} \theta_1^* &= \underset{\theta_1}{\text{Max}} F_c(\theta_1) \\ \theta_2^* &= t(\theta_1^*) \end{aligned}$$